基于扩展规则与统计特征的未登录词识别

曾　浩[1,2]，詹恩奇[1,2]，郑建彬[1,2]，汪　阳[1,2†]

(1. 武汉理工大学 信息工程学院，武汉 430070; 2. 光纤传感技术与信息处理教育部重点实验室，武汉 430070)

摘　要：为提高各行业领域未登录词识别效果，提出一种基于扩展规则与统计特征的未登录词识别方法。分析行业领域未登录词构词特点，制定扩展规则，根据扩展规则对分词项进行扩展得到复合词，通过词频、互信息、邻接熵等统计特征判别复合词是否为未登录词，若为未登录词，则对其继续扩展和识别。6 个行业领域和通用领域未登录词识别实验结果表明，提出方法取得了较好的未登录词识别效果，具有较好的移植性。

关键词：未登录词；扩展规则；词频；互信息；邻接熵

中图分类号：TP391　　doi: 10.3969/j.issn.1001-3695.2018.02.0140

Unregistered word recognition based on expansion rules and statistical features

Zeng Hao[1,2], Zhan Enqi[1,2], Zheng Jianbin[1,2], Wang Yang[1,2†]

(1. *College of Information Engineering*, *Wuhan University of Technology*, *Wuhan* 430070, *China*; 2. *Key Laboratory of Fiber Optic Sensing Technology &Information Processing of Ministry of Education*, *Wuhan* 430070, *China*)

Abstract: In order to improve unregistered word recognition effect in various fields, this paper proposed an unregistered word recognition method based on expansion rules and statistical features. It analyzed word formation features of unregistered words in various field, formulated expansion rules, extended word segmentations to get compound words according to expansion rules, then determined whether compound words were unregistered words through statistical features such as word frequency, mutual information and branch entropy, if the compound word was an unregistered word, it would continue to be expanded and recognized. The results of unregistered word recognition experiments in six fields and general field show that the method based on expansion rules and statistical features achieves better recognition effect of unregistered words and has better portability.

Key words: unregistered word; expansion rules; word frequency; mutual information; branch entropy

0　引言

在英语等西方语言书面表达中，句与句之间以标点符号为分隔符，词与词之间以空格为分隔符，计算机处理这些语言文本时，通过标点符号可以识别句子，通过空格可以识别词。在汉语书面表达中，虽然也以标点符号作为句子分隔符，但是词与词之间却无明显分隔符，字与字紧密相连，任何相邻的字都可能组成词，词的长度也没有限制。因此，计算机处理中文文本时，中文分词便成为了一项非常重要的基础工作。目前投入使用的各大分词器在通用领域取得了较高的分词准确率，但是其它行业领域分词效果并不理想。原因在于，各行业领域未登录词通常为长度更长语义更完整的复合词以及含有特殊字符的复合词，识别难度大。各行业领域未登录词识别准确率不高，就难以提高各行业领域文本分词准确率。因此，本文研究行业领域未登录词识别。

1　相关研究

未登录词指未被分词词典收录的词语以及随着时代发展而涌现出来的新词。其识别方法可分为基于规则的方法、基于统计的方法、规则与统计相结合的方法。

基于规则的方法通过构词模式、词性规则、成词概率等识别未登录词。郑家恒等人[1]研究汉语构词法，建立构词规则识别网络新词，取得了 91.2%的准确率。崔世起等人[2]通过语料库建立垃圾词典和词缀词典，结合词性规则和独立成词概率检测网络新词，也取得了较好的识别效果。基于规则的方法识别精度较高，但规则通常来源于特定领域，移植性较差，而且规则也不能概括所有的构词现象。

基于统计的方法认为词作为一个独立的整体，应具备稳定的内部结构和丰富的上下文环境，通常以词频、互信息、邻接熵等统计特征识别未登录词。韩艳等人[3]以互信息提取二元组，

收稿日期：2018-02-10；修回日期：2018-05-02

作者简介：曾浩（1994-），男，湖北武人，硕士研究生，主要研究方向为自然语言处理；詹恩奇（1972-），男，副教授，博士，主要研究方向为信号处理、模式识别；郑建彬（1966-），男，教授，博士，主要研究方向为模式识别、嵌入式系统；汪阳（1977-），男（通信作者），副教授，博士，主要研究方向为机器人控制、嵌入式系统（powerflow@whut.edu.cn）.

以邻接熵判断二元组边界并对其不断扩展，识别长度更长语义更完整的未登录词。杨阳等人[4]综合考虑词频、互信息、邻接熵等统计特征，提取长度不超过 6 的字符串，选取统计特征值均大于阈值的字符串为未登录词。李文坤等人[5]通过互信息从分词散串中筛选具有稳定结构的二元组，再对二元组进行扩展，以邻接熵判断词边界，识别长度为 2-4 的未登录词。夭荣朋等人[6]以改进互信息获取具有稳定结构的 2-gram 和 3-gram，通过计算 2-gram 的邻接熵以及对 3-gram 进行扩展识别未登录词。段宇锋等人[7]通过词频、文档频率、平均词频筛选一定范围内的候选项得到未登录词。Pang 等人[8]分析词在文档间、文档内、段落内的分布特征识别未登录词。Zhang 等人[9]以 K-means 方法对微博聚类，从每一类微博中提取词频大于阈值的候选串，通过邻接度判别候选串的子串是否为未登录词。基于统计的方法不依赖于规则，移植性较好，但计算量大，且由于没有规则的约束，结果中含有大量非词字符串。

为克服规则方法和统计方法的缺点，学者们更倾向于采用规则与统计相结合的方法，提高未登录词识别效果。Liu 等人[10]通过统计方法、领域词典、词性规则、前后缀规则等识别未登录词；霍帅等人[11]结合词频和词法规则识别未登录词；周超等人[12]综合词频、词性规则和邻接变化数识别未登录词；杜丽萍等人[13]以改进互信息筛选二元组并对其扩展，通过词频规则和停用词规则过滤得到未登录词。

大多数研究以新闻、微博为语料，研究通用领域未登录词识别方法，识别对象主要是长度为 2-4 的中文未登录词，缺乏对长度更长语义更完整的复合词的识别研究。此外，对中文文本中含英文的特殊未登录词识别研究相对较少。

本文研究行业领域未登录词识别，提出一种基于扩展规则与统计特征的未登录词识别方法。以 6 个行业领域招聘职位为语料，分析行业领域未登录词构词特点，建立扩展规则，根据扩展规则对分词项扩展得到复合词，综合词频、互信息、邻接熵等统计特征判别复合词是否为未登录词，若为未登录词，则继续扩展和识别。

2　行业领域未登录词识别

通过网络爬虫从招聘网站爬取招聘职位，建立职位语料库，招聘职位如图 1 所示，职位语料库如表 1 所示。职位通常由两部分组成：结构化数据和非结构化数据。结构化数据包括职位月薪、工作地点、发布时间等字段及相应内容，这部分内容通常由若干字描述。非结构化数据包括岗位职责、任职要求、福利待遇等。职位信息主要集中在非结构化数据，因此在本文后续工作中，关于职位的处理指的是对其非结构化数据的处理。

从每个行业领域各提取 50 个职位，使用分词器 HanLP 进行分词。造成分词错误的主要因素是歧义和未登录词，因此，排除分词结果中由歧义造成的分词错误字段，剩下的分词错误字段便可认为是未登录词。6 个行业领域未登录词统计如表 2 所示。

职位月薪：6000-12000元/月　　工作地点：北京-昌平区
发布日期：2017-10-10　　工作性质：全职
工作经验：无经验　　最低学历：本科
招聘人数：10人　　职位类别：算法工程师

岗位职责：
1.参与大数据或云架构或机器学习相关的开发；
2.参与机器学习和人工智能的设计和开发实现；
3.参与数学建模、数据挖掘、自然语言处理等相关模块研究；
任职要求：
1.统招本科及以上学历，数学、计算机科学、通信工程等相关专业
2.熟悉C/C++/Java等开发语言，熟悉常用算法和数据结构；
3.熟悉机器学习、深度学习，使用过相关开源框架者优先；
福利待遇：
1、基本工资+五险一金+餐补+法定假日+带薪年假+节日福利+生日福利+年度旅游。
2、朝九晚五，五天八小时工作制，周末双休！
工作地址：
北京市昌平区北七家镇未来科技城南区中国电子信息安全技术研发基地B栋5层

图 1　招聘职位

表 1 职位语料库

行业领域	职位个数
IT 互联网	83753
财务/人力/行政	40189
销售/客服/市场	83401
项目/质量/管理	32788
房产/建筑/物业	35093
金融	39376

表 2 6 个行业领域未登录词统计

行业领域	职位个数	未登录词个数
IT 互联网	50	346
财务/人力/行政	50	319
销售/客服/市场	50	325
项目/质量/管理	50	275
房产/建筑/物业	50	302
金融	50	342

表 2 中未登录词可分为中文未登录词和英文未登录词。其中，中文未登录词约占 90%，主要为人名、地名、机构名、行业术语。英文未登录词约占 10%，主要为表示工作技能的行业术语，如“c++”、“j2se”。对中文未登录词和英文未登录词的构词特点分析，如表 3、4 所示。

表 3 中文未登录词构词特点

特点	实例	比例(%)
1+1	餐补，入职，电销，调优，直招	10.00
1+2	微商城，云产品，大数据，高并发	9.00
2+1	工龄奖，通讯费，招商部，季度奖	10.00
2+2	深度学习，淘宝客服，市场营销	35.00
2+3	通信运营商，注册会计师，办公自动化	15.00
3+2	新媒体运营，房地产开发，节假日福利	8.00
2+2+2	自然语言处理，语音信号处理	10.00
其它	计算机科学与技术，电子与通信工程	3.00

表 4 英文未登录词构词特点

特点	实例	比例(%)
英文+中文	c 语言，ip 协议	10.00
英文+数字	html5，spring3，stm32	50.00

英文+特殊字符	c++，c#，notepad++	5.00
英文+数字+英文	j2se，j2ee，p2p	10.00
英文+特殊字符+英文	asp.net，object-c，b/s	20.00
英文+特殊字符+数字	cet-4，cet-6	5.00

由表 3 可知，对于各行业领域中文未登录词，其一般是由 2-3 个中文词组成的复合词。由表 4 可知，英文未登录词通常也是由 2-3 部分组成的复合词，但它的构词特点比中文词更灵活。中文词通常只和中文词组成复合词，而英文词既可以与中文词组成复合词，如“c 语言”，也可以和数字组成复合词，如“html5”，甚至还可以和特殊字符组成复合词，如“c#”。HanLP 因未能识别这些未登录词，将它们错误切分为若干个分词项。例如，将“深度学习”错误切分为“深度/学习”，将“j2ee”错误切分为“j/2/ee”。因此，若能根据未登录词构词特点，将分词结果中的分词项按照一定规则进行重组，再通过某种策略过滤，便可识别各行业领域未登录词。

3　基于扩展规则与统计特征的未登录词识别

3.1　扩展规则

在分析行业领域未登录词构词特点的基础上，提出基于扩展规则与统计特征的未登录词识别方法。方法中的扩展指：HanLP 分词后，同一句分词结果中当前词与后一个词组成复合词。扩展规则具体如下：

Rule1　当前词为停用词或者既不是中文词也不是英文词，则当前词不扩展。

Rule2　当前词为中文词且不是停用词，如果后一个词也为中文词且不是停用词，则当前词扩展。

Rule3　当前词为英文词且不是停用词，如果后一个词不是停用词，则当前词扩展。

Rule4　扩展次数大于预设最大扩展次数，不再扩展。

上述扩展规则源于对行业领域未登录词构词特点的总结。中文词通常只和中文词组成复合词，而英文词则可以与中文词、数字、特殊字符等组成有意义的复合词。因此，上述扩展规则既可筛选符合行业领域未登录词构词特点的复合词，又可去除一些无意义的组合，提高未登录词识别效果。

扩展规则需要使用停用词词典，在自然语言处理中，停用词指只在语句中充当某种成分而对语义表达无任何贡献的字词，这些字词通常不与其它字词构成有意义的复合词，比如“了”、“的”、“不”。互联网上存在各种版本的停用词词典，这些停用词词典通常只收录通用领域的停用词。而本文研究涉及各行业领域，为提高各行业领域未登录词识别效果，对职位语料库进行分词并统计词频，从中选取词频大于 1000 且与其它词组成复合词概率低的词作为停用词，部分停用词及其词频如表 5 所示。再结合通用领域停用词词典，建立一部含 1900 个停用词的行业领域停用词词典。

表 5　停用词及其词频

停用词	词频	停用词	词频	停用词	词频
相关	246481	熟练	80161	参与	57928
以上	207463	及时	74945	安排	54905
具有	189815	使用	72890	各项	54370
进行	185871	了解	70874	各种	54130
熟悉	176847	其他	69804	做好	52225
完成	168965	能够	68439	建立	45268
提供	142157	以及	63221	各类	43410
具备	126778	我们	61132	善于	35422

3.2　统计特征

本文以词频、互信息、邻接熵作为未登录词识别的统计特征。如果扩展所得的复合词的统计特征值均大于阈值，则判定为未登录词，否则不是未登录词。

1）词频

未登录词作为词，首先应具备一定的出现次数。记 $f(w)$ 表示复合词 w 在语料库中出现的次数，$f(w)$ 越大，复合词 w 成为未登录词的可能性越大。

2）互信息

未登录词作为词，应具备稳定的内部结构。信息论中，互信息(mutual information, MI)用于衡量两个信号的关联程度。因此，互信息也可衡量两个词结合的紧密程度。互信息越大，结合得越紧密，相邻词组成的复合词成为未登录词的概率越大。互信息计算公式如式（1）～（4）所示。

$$MI(w)=\log(\frac{p(w)}{p(x)p(y)}) \quad (1)$$

$$p(w)=\frac{f(w)}{N} \quad (2)$$

$$p(x)=\frac{f(x)}{N} \quad (3)$$

$$p(y)=\frac{f(y)}{N} \quad (4)$$

其中：w 表示由词 x 和词 y 组成的复合词，$MI(w)$ 表示 w 的互信息，$p(w)$、$p(x)$、$p(y)$ 分别表示 w、x、y 在语料库中出现的概率，$f(w)$、$f(x)$、$f(y)$ 分别表示 w、x、y 在语料库中的词频，N 表示语料库中的总词数。

式 (1) 只适用于计算由两个词组成的复合词的互信息，为计算由多个词组成的复合词的互信息，对式 (1) 进行改进，改进后的互信息如式（5）所示 。

$$MMI(w)=\log(\frac{p(w)}{Avg(w_1...w_n)}) \quad (5)$$

$$Avg(w_1...w_n)=\frac{1}{n-1}\sum_{i=1}^{n-1}p(w_1...w_i)p(w_{i+1}...w_n) \quad (6)$$

其中：w_1，…，w_i，…，w_n 为组成复合词 w 的 n 个词，$MMI(w)$ 为改进后复合词 w 的互信息，$Avg(w_1w_2...w_n)$ 为组成复合词 w 的不同组合的平均概率。例如，对于由“自然”“语言”“处理”三个词组成的复合词“自然语言处理”，组

成它的不同组合的平均概率为：P(自然)P(语言处理) 和 P(自然语言)P(处理) 均值。

3）邻接熵

未登录词作为词，应具备丰富的上下文环境。邻接熵(branch entropy, BE)是衡量字符串成词概率的一个重要统计特征，利用信息熵(information entropy, IE)计算字符串上下文的不确定性[14]。信息论中，信息熵用于表示随机变量的不确定性均值，随机变量的信息熵越大，它的不确定性就越大。假设 A 是一个离散型随机变量，取值空间为 B，当 A 取值 $a \in B$ 时，概率分布为 $P(a) = P(A = a)$，随机变量 A 的信息熵如式（7）所示。

$$IE(A) = -\sum_{a \in B} p(a)\log p(a) \qquad (7)$$

邻接熵分为左邻接熵(left branch entropy, LBE)和右邻接熵(right branch entropy, RBE)。如果字符串的左邻接熵越大，其上文环境越丰富，那么它的左边界就可以确定；如果字符串的右邻接熵越大，其下文环境越丰富，那么它的右边界就可以确定；如果字符串的左邻接熵和右邻接熵均很大，其左右边界均可确定，那么它单独成词概率就越大。

本文中，复合词的左邻接指它的前一个词，右邻接指它的后一个词，所有左邻接构成左邻接集合，所有右邻接构成右邻接集合，所有不同的左邻接构成左邻接类别，所有不同的右邻接构成右邻接类别。假设复合词 w 的左邻接类别为 $LS(w) = \{L_1, L_2, ..., L_i, ..., L_n\}$，右邻接类别为 $RS(w) = \{R_1, R_2, ..., R_i, ..., R_m\}$，其左邻接熵和右邻接熵分别如式（8）（9）所示。

$$LBE(w) - \sum_{i=1}^{n} \frac{n_i}{n} \log \frac{n_i}{n} \qquad (8)$$

$$RBE(w) = -\sum_{i=1}^{m} \frac{m_i}{m} \log \frac{m_i}{m} \qquad (9)$$

式 (8) 中，$LBE(w)$ 表示 w 的左邻接熵，n 表示左邻接集合的大小，n_i 表示左邻接集合中左邻接 L_i 出现的次数。式 (9) 中，$RBE(w)$ 表示 w 的右邻接熵，m 表示右邻接集合的大小，m_i 表示右邻接集合中右邻接 R_i 出现的次数。

3.3　未登录词识别流程

基于扩展规则与统计特征的未登录词识别流程如图 2 所示。具体步骤如下：

a)设置最大扩展次数、词频阈值、互信息阈值、左邻接熵阈值、右邻接熵阈值。

b)将语料库按中文标点符号切分为短句。

c)使用 HanLP 对短句分词，遍历分词项，根据扩展规则判断当前词是否可扩展。如果当前词不可扩展，则跳过此当前词，并将后一个分词项作为当前词进行扩展和识别。如果当前词可扩展，计算扩展所得的复合词的统计特征值，若均大于阈值，则添加到未登录词集合，并对此复合词继续扩展和识别；否则，舍弃该复合词，并将后一个分词项作为当前词进行扩展和识别。

d)语料库中所有短句处理完毕，算法结束，输出未登录词集合。

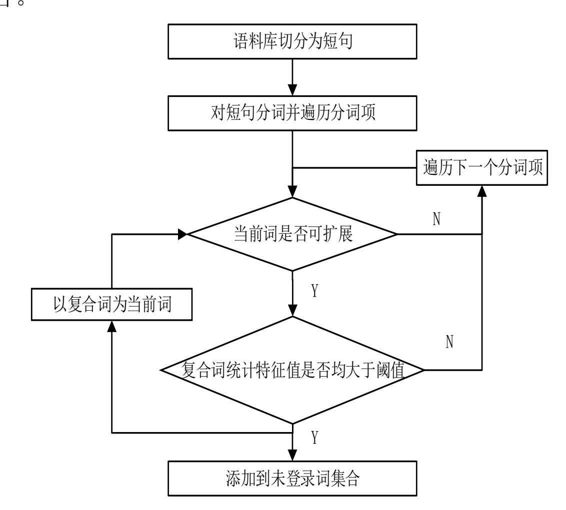


图 2　基于扩展规则与统计特征的未登录词识别流程

4　实验及分析

4.1　实验方法及评价标准

本文方法参数设置：最大扩展次数为 2，即只识别由两个词和三个词组成的复合词。词频阈值为 10；互信息阈值为 3；左邻接熵阈值和右邻接熵阈值均为 1。考虑到语料库中低频未登录词，本文方法将各统计特征阈值设置较低，尽可能识别出更多未登录词。

文献[4]提取长度为 2~6 的字符串，若字符串的词频、互信息、左邻接熵、右邻接熵均大于阈值，则判定为未登录词。而本文方法是识别长度更长语义更完整的复合词，为了与本文方法对比，对文献[4]方法稍加修改，不再提取长度为 2~6 的字符串，而是提取由相邻分词项组成的二元组和三元组，若二元组和三元组的词频、互信息、左邻接熵、右邻接熵均大于阈值，则判定为未登录词。此外，各参数值均与本文方法参数值相同。

文献[5]在分词的基础上从散串中提取互信息和词频均大于阈值的二字组合，然后通过左邻接熵和右邻接熵对二字组合进行扩展，主要识别长度为 2-4 的未登录词。而本文方法是识别长度更长语义更完整的复合词，为了与本文方法对比，对文献[5]方法稍加修改，不再从散串中提取二字组合，而是提取由相邻两个分词项组成的二元组，后续的扩展和识别保持和文献[5]一样。此外，各参数值均与本文方法参数值相同。

以准确率(P)、召回率(R)和 F 值(F)作为未登录词识别结果的评价标准，如式 (10)～(12)。

$$P = \frac{|C \cap D|}{|C|} \times 100\% \qquad (10)$$

$$R = \frac{|C \cap D|}{|D|} \times 100\% \qquad (11)$$

$$F = \frac{2PR}{P + R} \qquad (12)$$

其中：C 表示方法识别出的未登录词集合，D 表示人工标注的未登录词集合。

4.2　行业领域未登录词识别对比

运用文献[4,5]、本文方法，识别表 2 中 6 个行业领域各 50 个职位中未登录词，实验结果如表 6~8 所示。

表 6　行业领域未登录词识别准确率(%)对比

行业领域	文献[4]	文献[5]	本文方法
IT 互联网	57.14	54.38	60.26
财务/人力/行政	56.41	55.28	58.36
销售/客服/市场	54.55	51.92	60.66
项目/质量/管理	51.11	51.33	59.74
房产/建筑/物业	42.17	42.98	52.08
金融	53.14	53.50	54.91

表 7　行业领域未登录词识别召回率(%)对比

行业领域	文献[4]	文献[5]	本文方法
IT 互联网	32.37	34.10	52.60
财务/人力/行政	27.59	27.90	51.41
销售/客服/市场	31.38	33.23	56.92
项目/质量/高级	25.10	28.00	51.27
房产/建筑/物业	32.12	34.44	53.97
金融	37.13	38.01	60.53

表 8　行业领域未登录词识别 F 值(%)对比

行业领域	文献[4]	文献[5]	本文方法
IT 互联网	41.33	42.00	56.17
财务/人力/行政	37.05	37.10	54.67
销售/客服/市场	39.84	41.00	58.73
项目/质量/高级	33.66	36.24	55.19
房产/建筑/物业	36.47	38.24	53.01
金融	43.72	44.45	57.58

实验结果表明，本文方法在识别行业领域未登录词中取得了较好效果，其准确率、召回率、F 值均高于另外两种方法。文献[4,5]均以微博为语料库，研究通用领域未登录词识别，虽然充分利用了词频、互信息、邻接熵等统计特征，但是缺少对规则的运用，未登录词识别结果中包含大量统计特征值大于阈值的非词字符串。例如，文献[4]方法在识别 IT 互联网行业未登录词中，识别结果包含“学习 Java”，这是因为“学习”和“Java”在此行业语料中共现次数较高，导致“学习 Java”具有较高的词频、互信息、邻接熵。本文方法不仅充分利用了词频、互信息、邻接熵等统计特征，同时还结合了扩展规则，扩展规则源于对各行业领域未登录词构词特点的总结，中文词通常只和中文词组合成复合词，而英文词可以和中文词、数字、特殊符号等组合成有意义的复合词。根据扩展规则，可以避免类似“学习 Java”这样无意义组合的产生，在一定程度上提高了未登录词识别效果。

4.3　通用领域未登录词识别对比

微博覆盖内容较广，属于通用领域数据。从 COAE2014 提供的数据中选取 5 000 条微博作为实验数据，分别采用文献[4,5]、本文方法识别其中的未登录词。由于难以标注微博中未登录词，故仅以准确率作为实验结果的评价标准，实验结果如表 9 所示。

表 9　通用领域未登录词识别准确率(%)对比

方法	识别个数	正确个数	准确率
文献[4]	404	254	62.87
文献[5]	496	304	61.29
本文方法	469	336	71.64

实验结果表明，本文方法在识别微博未登录词中取得了较好效果，其准确率高于文献[4,5]方法。

5　结束语

本文对行业领域未登录词识别方法进行研究，通过网络爬虫技术爬取各行业领域招聘职位，在分析行业领域未登录词构词特点的基础上，制定扩展规则，根据扩展规则对分词项进行扩展得到复合词，再综合词频、互信息、邻接熵等统计特征判定复合词是否为未登录词。在 6 个行业领域以及通用领域进行未登录词识别实验，本文方法取得了较好的准确率、召回率和 F 值，说明本文方法是有效的，具有较好的移植性。由于本文方法根据统计特征值是否均大于阈值来判断复合词是否为未登录词，判断条件过于苛刻，因此无法识别出部分统计特征值大于阈值的未登录词，例如，未能识别出词频、互信息和左邻接熵均大于阈值但右邻接熵低于阈值的未登录词。今后将针对这一问题进行改进，进一步提高未登录词识别效果。

参考文献：

[1] 郑家恒，李文花．基于构词法的网络新词自动识别初探 [J]．山西大学学报：自然科学版, 2002, 25 (2): 115-119. (Zheng Jiaheng, Li Wenhua. A new approach to automatic recognition of web new words based on word formation [J]. Journal of Shanxi University: Natural Science Edition, 2002, 25 (2): 115-119.)

[2] 崔世起，刘群，孟遥，等．基于大规模语料库的新词检测 [J]．计算机研究与发展, 2006, 43 (5): 927-932. (Cui Shiqi, Liu Qun, Meng Yao, etal. New word detection based on large-scale corpus [J]. Journal of Computer Research and Development, 2006, 43 (5): 927-932.)

[3] 韩艳，林煜熙，姚建民．基于统计信息的未登录词的扩展识别方法 [J]．中文信息学报, 2009, 23 (03): 24-30, 50. (Han Yan, Lin Yuxi, Yao Jianmin. Extended identification method for unregistered words based on statistical information [J]. Journal of Chinese Information Processing, 2009, 23 (03): 24-30, 50.)

[4] 杨阳，刘龙飞，魏现辉．基于词向量的情感新词发现方法 [J]．山东大学学报：理学版, 2014, 49 (11): 51-58. (Yang Yang, Liu Longfei, Wei Xianhui. Emotional new word discovery method based on word vector [J]. Journal of Shandong University: Science Edition, 2014, 49 (11): 51-58.)

[5] 李文坤，张仰森，陈若愚. 基于词内部结合度和边界自由度的新词发现 [J]. 计算机应用研究, 2015, 32 (8): 2302-2304, 2342. (Li Wenkun, Zhang Yangsen, Chen Ruoyu. New word discovery based on internal conjunction degree and boundary freedom [J]. Application Research of Computers, 2015, 32 (8): 2302-2304, 2342.)

[6] 夭荣朋，许国艳，宋健. 基于改进互信息和邻接熵的微博新词发现方法 [J]. 计算机应用, 2016, 36 (10): 2772-2776. (Yao Rongpeng, Xu Guoyan, Song Jian. Microblog new word discovery method based on improved mutual information and adjacency entropy [J]. Journal of Computer Applications, 2016, 36 (10): 2772-2776)

[7] 段宇锋，鞠菲. 基于 N-Gram 的专业领域中文新词识别研究 [J]. 数据分析与知识发现, 2012, 28 (2): 41-47. (Duan Yufeng, Ju Fei. Research on recognition of chinese new words in professional field based on N-Gram [J]. Data Analysis and Knowledge Discovery, 2012, 28 (2): 41-47.)

[8] Pang Wenbo, Fan Xiaozhong, Gu Yijun, *et al.* Chinese unknown words extraction based on word-level characteristics [C]// Proc of International Conference on Hybrid Intelligent Systems. 2009: 361-366.

[9] Zhang Shuai, Liu Qianren, Wang Lei. A Weibo-oriented method for unknown word extraction [C]// Proc of the 8th International Conference on Semantics, Knowledge and Grids. Washington DC: IEEE Computer Society, 2012: 209-212.

[10] Liu Qingtang, Wu Linjing, Yang Zongkai, etal. Domain phrase identification using atomic word formation in Chinese text [J]. Knowledge-Based Systems, 2011, 24 (8): 1254-1260.

[11] 霍帅，张敏，刘奕群. 基于微博内容的新词发现方法 [J]. 模式识别与人工智能, 201427 (02): 141-145. (Huo Shuai, Zhang Min, Liu Yiqun. New word discovery method based on weibo content [J]. Pattern Recognition and Artificial Intelligence, 2014, 27 (2): 141-145.)

[12] 周超，严馨，余正涛. 融合词频特性及邻接变化数的微博新词识别 [J]. 山东大学学报：理学版, 2015, 50 (3): 6-10. (Zhou Chao, Yan Xin, Yu Zhengtao. Micro-blog new word recognition based on word frequency feature and adjacency change number [J]. Journal of Shandong University: Science Edition, 2015, 50 (3): 6-10.)

[13] 杜丽萍，李晓戈，于根. 基于互信息改进算法的新词发现对中文分词系统改进 [J]. 北京大学学报：自然科学版, 2016, 52 (1): 35-40. (Du Liping, Li Xiaoge, Yu Gen. The improvement of chinese word segmentation based on new word discovery based on improved mutual information [J]. Journal of Peking University: Natural Science Edition, 2016, 52 (1): 35-40.)

[14] Zhikov V, Takamura H, Okumura M. An efficient algorithm for unsupervised word segmentation with branching entropy and MDL [J]. Transactions of the Japanese Society for Artificial Intelligence, 2013, 28 (3): 347-360.